جامعة عبد المالك السعدي
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵏ ⵄⴰⴱⴷⵍⵎⴰⵍⵉⴽ ⵙⵙⴰⵄⴷⵉ
Université Abdelmalek Essaadi

Pôle des Etudes Doctorales
Centre des Etudes Doctorales Sciences et Techniques et Sciences Médicales
Formation Doctorale : STI
Etablissement : ENSA de TANGER

Nom et Prénom : Mohamed EDDOUJAJI
Date de la soutenance : 5 Octobre 2024
Directeur de Thèse : Hassan SAMADI
Discipline : STI
Spécialité : Informatique
Structure de recherche : *LABTIC*
Intitulé de la thèse : MAXIMISATION DE L'EFFICACITE DES PROCESSUS MAP REDUCE DANS UN ÉCOSYSTEME HADOOP PAR L'OPTIMISATION DES ALGORITHMES DE LA DATA LOCALITY, PLANIFICATION ET PRELECTURE

Résumé

Au cours des deux dernières décennies, l'émergence des nouvelles technologies a entraîné une explosion du volume de données à gérer, communément appelé le "Big Data", engendrant ainsi des défis significatifs en matière de stockage et de traitement de ces données. C'est dans ce contexte que le framework logiciel Hadoop se distingue en offrant des solutions adaptées à ces contraintes.

L'objectif de cette thèse est d'explorer à bon escient les éléments de base de Hadoop (HDFS, MapReduce et YARN) dans le but d'améliorer et d'optimiser les performances globales d'un tel système. Le travail est scindé en trois parties :
Dans la première partie, nous nous sommes intéressés à l'impact des petits fichiers dans l'environnement Hadoop et leur défis associés à la gestion et au traitement efficace dans un contexte distribué. Nous avons réussi à mettre en place un nouvel algorithme de fusion (Merging Algorithm) qui consiste à fusionner les petits fichiers en fonction de leur volume, contrôlé par un seuil de mémoire tampon configurable. Cette nouvelle approche nous a permis d'améliorer les performances et l'efficacité du traitement au sein de l'écosystème Hadoop (consommation de la mémoire, la durée de stockage et de lecture).
La deuxième partie de ce travail est consacrée au placement stratégique des données dans un environnement distribué. Nous avons analysé les différentes stratégies de placement des données et évaluer leur impact sur les performances globales du système. Nous avons mis en place une nouvelle politique de placement des données (Algorithme DPP) qui prend en considération la différence entre les capacités de calcul des nœuds du cluster. Nous avons ainsi optimisé le positionnement des données pour une utilisation plus efficace des ressources et une amélioration des temps de traitement.
Enfin, dans la troisième partie, en utilisant les techniques de prélecture et de planification de Hadoop pour anticiper les besoins en données, nous avons mis en place un framework qui propose un nouveau mécanisme de planification et de prélecture en considérant les modèles d'accès aux données comme un facteur crucial pour améliorer non seulement la localité des données, mais aussi l'utilisation de métriques pour prendre de meilleures décisions lors de la réclamation d'allocation de ressources.

Dans chaque partie, des tests de simulation ont été effectués pour valider tous les résultats obtenus.

Mots clés: Big Data, Distributed Systems, Heterogeneous Systems, Hadoop Distributed File System; Distributed Storage; Distributed Computing, Sequencefile, scheduling, prefetching